UNIVERSITY OF EXETER
Q-STEP CENTRE

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE

Summary statistics with Python

Lewys Brace
l.brace@Exeter.ac.uk

Summary statistics

- Graphing data is a great first step in your data analysis.
- However, you are probably also going to want to produce a more succinct, numerical-based, summary of you data in order to compliment your graphical analysis.

What is a typical score?

- If we wanted to summarise the percentage of votes of Obama at the county level in Pennsylvania in one number, what would we choose?
- We need a measure of central tendency, options include:
    - The *mean,* also known as the average, or the arithmetic mean.
    - The *median*, or middle score in the range of data.
    - The *mode*, the most frequently occurring score.

The mean ($\bar{x}$)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Calculating the mean = $\frac{sum\ of\ all\ data\ values}{number\ of\ data\ values}$

- Let's imagine that we have a hypothetical distribution, consisting of 15 observations.
- This distribution could be anything, lets say is the number of fish I've caught during a day.
- The weight of these 15 fish are:
  2.1, 2.4, 2.4, 2.4, 2.4, 2.6, 2.9, 3.2, 3.2, 3.9, 4.5, 6.3, 8.2, 12.8, 23.5
- The sum of these fish weights is 82.8.
- There are 15 values, so:

$$\bar{x} = \frac{82.8}{15}$$
$$\bar{x} = 5.52$$

Mean vote percentage

- The mean of a variable can easily be calculated in Python using Numpy's mean() function.

In:

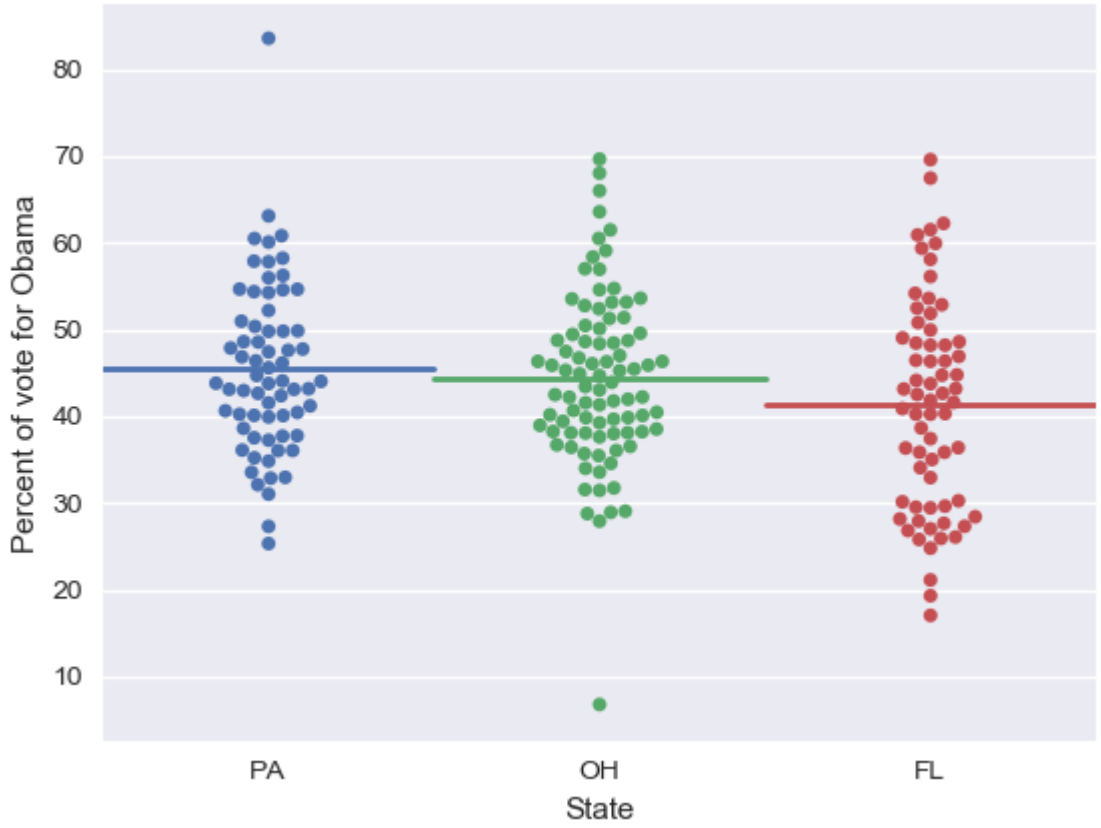
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
df_swing = pd.read_csv('C:\Users\lb690\Google Drive\Teach
df_PA = df_swing.loc[df_swing['state'] == 'PA']
PA_mean = np.mean(df_PA['dem_share'])
print "Pennsylvania democratic vote mean:", PA_mean
```

Line that calculates the mean.

Out:

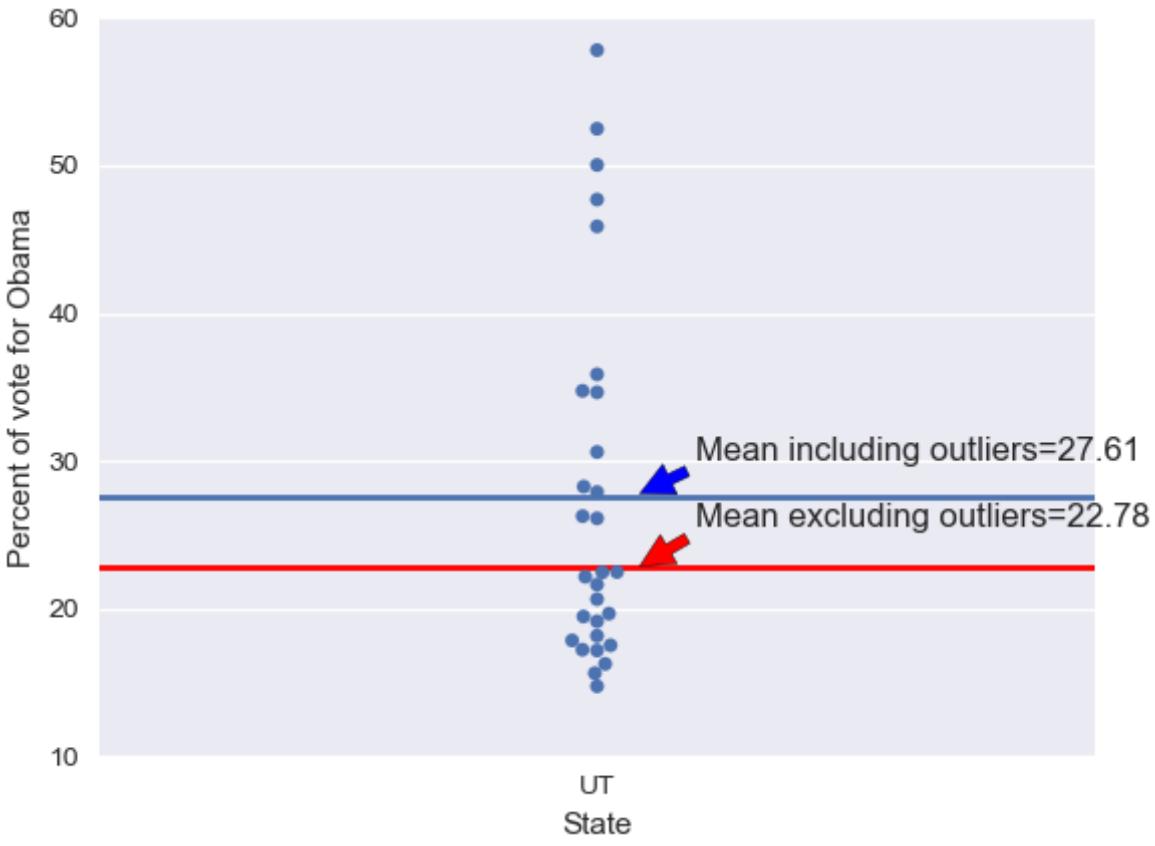
```
Pennsylvania democratic vote mean: 45.4764179104
```

- If we calculate the means for each state and add these to the plot as a line, we see that the mean is a reasonable summary of the data.
- However, the mean is highly susceptible to…



Outliers

- Outliers are data points whose value is far greater or far less than the rest of the data points.
- If we take a look at the county level data for Utah in the 2008 election, we see that there are 5 outliers.



- So, we may want a summary statistic that is immune to the impact of outliers...
- This is where the *median* comes in.
- Because the mean is calculated based on the rankings of the data and not on the value of the data, it is immune to outliers.

The median

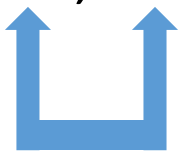
- To calculate the median, order all data values in regards to their values from smallest to largest. The median value is then the middle value in the range of scores.
- In our fish example:

2.1, 2.4, 2.4, 2.4, 2.4, 2.6, 2.9, 3.1, 3.2, 3.9, 4.5, 6.3, 8.2, 12.8, 23.5

median = 3.1

- If the data set has an even number of data values, you take the middle two numbers and then take the mean of these two numbers.
- If we add another data point into our fish example:

2.1, 2.4, 2.4, 2.4, 2.4, 2.6, 2.9, 3.1, 3.2, 3.9, 4.5, 6.3, 8.2, 12.8, 23.5, 23.6



$$\text{median} = \frac{3.1+3.2}{2}$$

$$\text{median} = \frac{6.3}{2}$$

$$\text{median} = 3.15$$

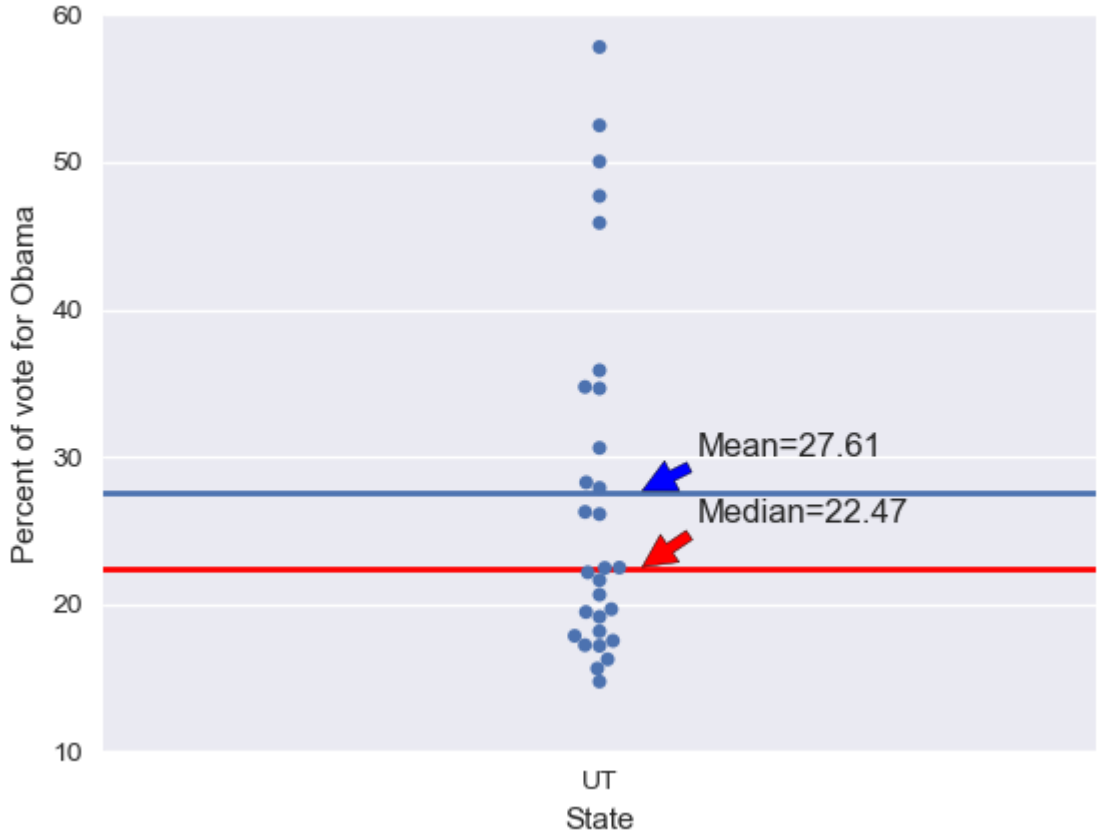
Calculating the median

- We can see here that the median is not ‘tugged up’ by the 5 outliers.
- Calculating the median in python is done in the same way as the mean:

In:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df_swing = pd.read_csv('C:\Users\lb690\Google Drive\Teachin
df_UT = df_swing.loc[df_swing['state'] == 'UT']
UT_median = np.median(df_UT['dem_share'])
print "Utah democratic vote median:", UT_median
```

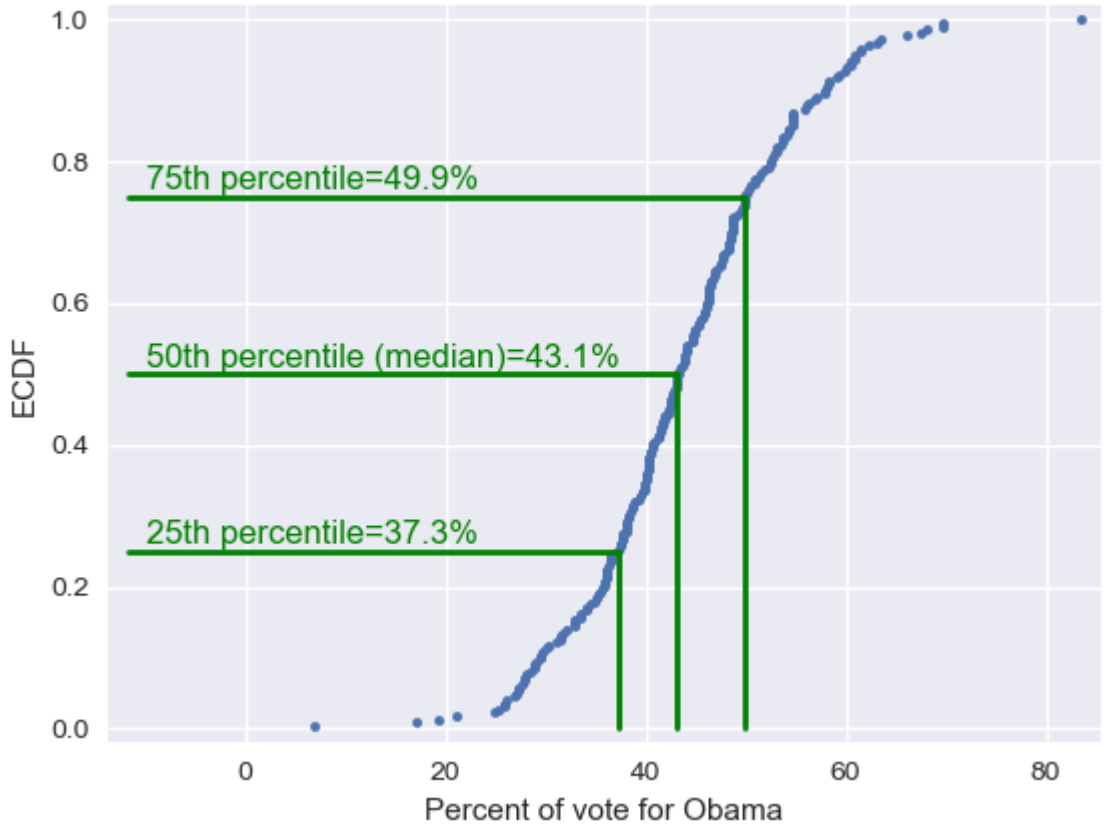
Line that calculates the mean.

Out:
```
Utah democratic vote median: 22.47
```



Percentiles

- The median is a special name for the 50th percentile.
- This means that 50% of the data are less than the median.
- Similarly, the 25th percentile is the data point that is equal to 25% of the data.
- And so on for 75th percentile, etc.



- Percentiles are useful summary statistics.
- They can easily be computed in Python with the help of Numpy:

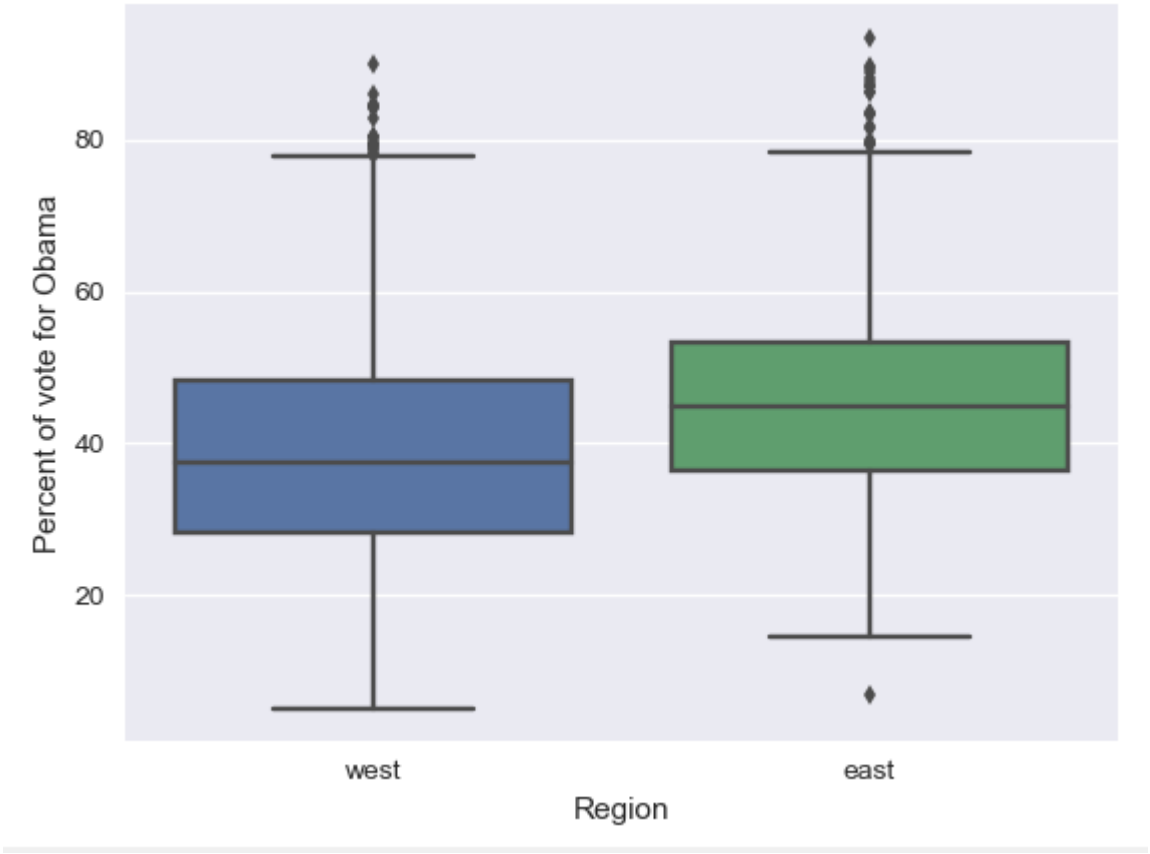
```
In:  import pandas as pd
     import matplotlib.pyplot as plt
     import numpy as np
     import seaborn as sns
     from decimal import *
     sns.set()
     df_swing = pd.read_csv('C:\Users\lb690\Google Drive\Teaching\Statistics\Data_a
     swing_state_percentiles = np.percentile(df_swing['dem_share'], [25, 50, 75])
```

```
Out: percentiles [37.3025 43.185  49.925 ]
```

- We now have three summary statistics.
- However, the point of summary statistics is to keep things concise, but we are starting to get a lot of different numbers here.
- This is where quantitative EDA meets graphical EDA.

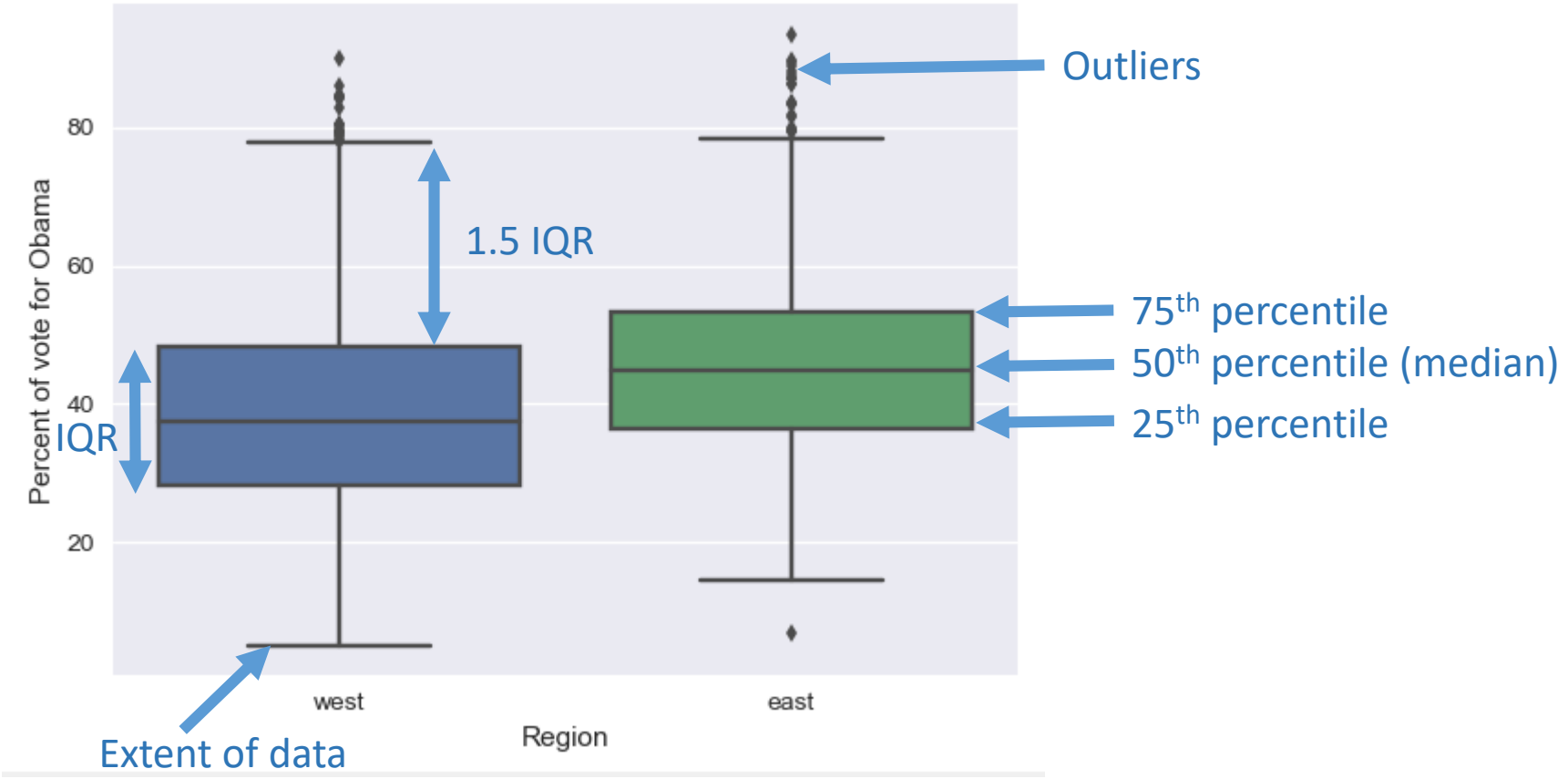
Box plots

- These were invented by John Tukey in order to display salient features of a dataset based on percentiles.
- Here, we have a box plot showing Obama's vote share for states east and west of the Mississippi River.



What does a box plot show?

- Interquartile range (IQR) = middle 50% of the data.
- The whiskers extend a distance of 1.5 times the IQR, or the extent of the data; whichever is less extreme.
- All point outside of the whiskers are plotted as individual points, which is the common criterion for defining an outlier.



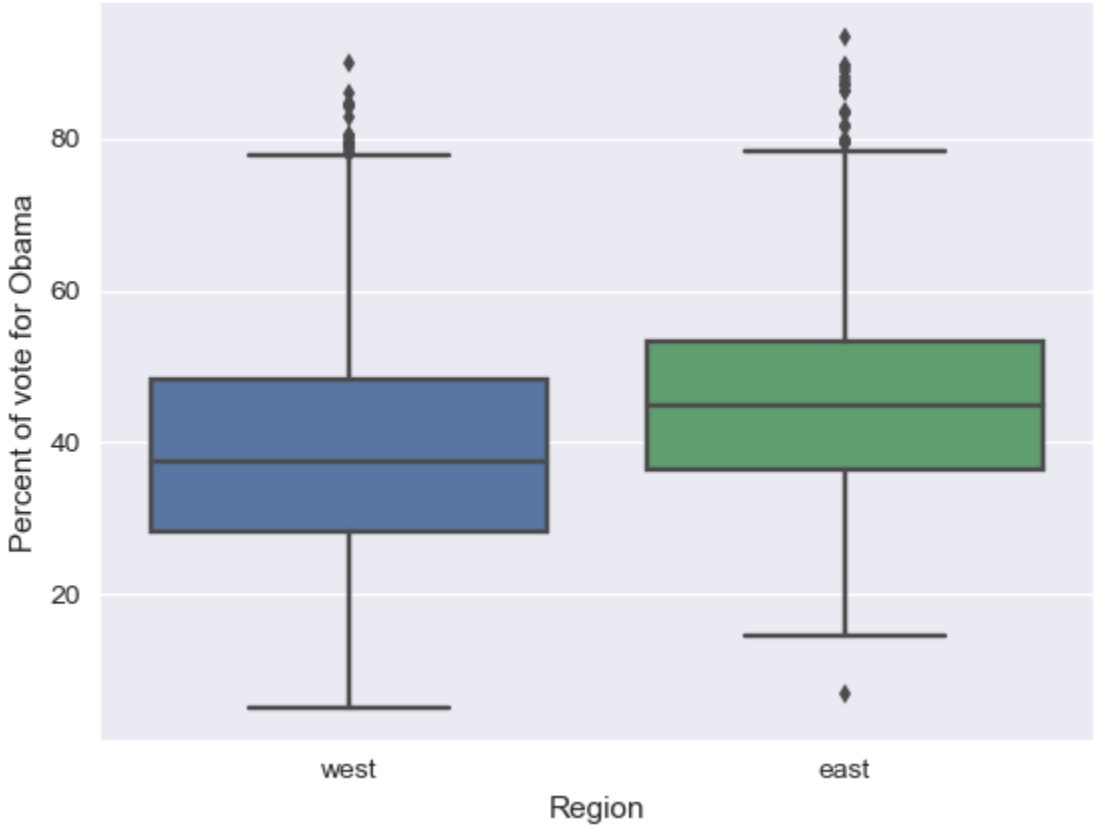
- When data are very large and bee swarm plots become too cluttered, box plots are a great alternative.
- It makes sense therefore that making a box plot in python is similar to making a bee swarm plot.

Plotting a box plot

In:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from decimal import *
sns.set()
df_all = pd.read_csv('C:\Users\lb690\Google Drive\Teaching\St
_=sns.boxplot(x='east_west', y='dem_share', data=df_all)
_=plt.xlabel('Region')
_=plt.ylabel('Percent of vote for Obama')
plt.show()
```

Out:

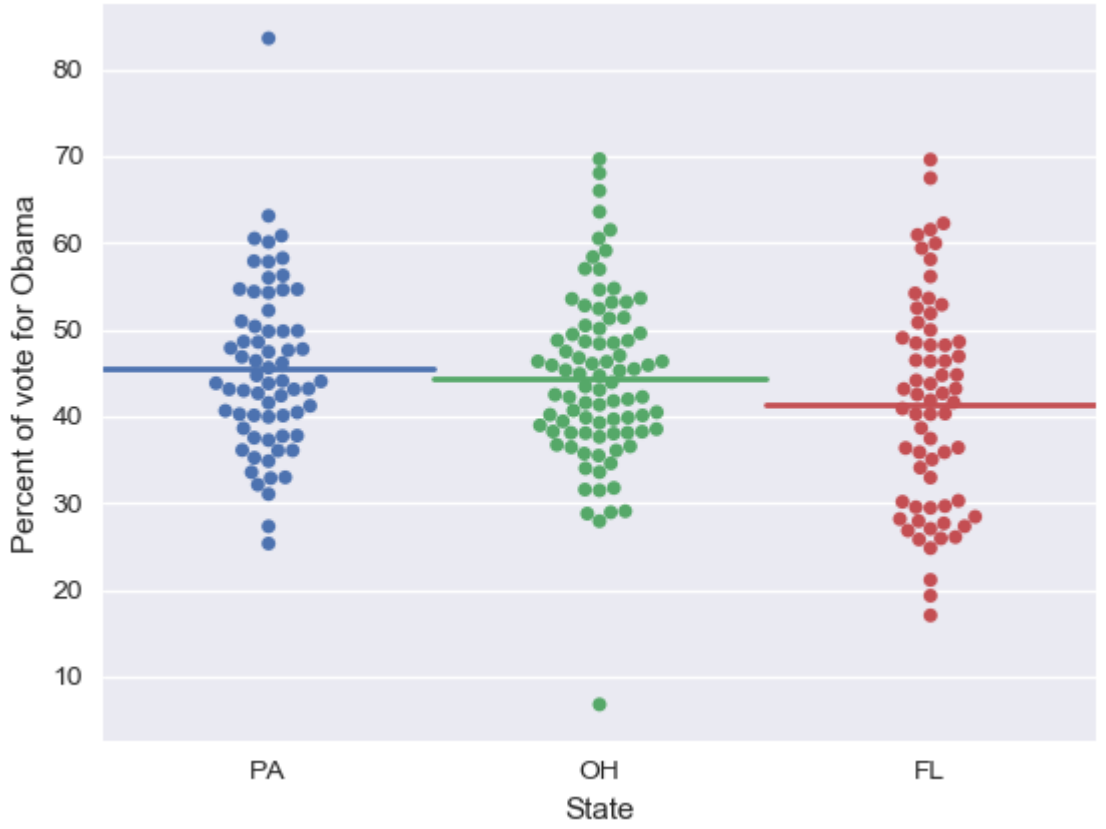


A note about outliers

- It is important to remember that an outlier is not necessarily an erroneous data point.
- You should not treat it as such unless you have a good reason to do so.
- Since there was a lack of any substantial voter fraud in the US, these outliers are not erroneous; they are just data point with extreme values.

Variance and Standard deviation

- Looking at the swing state data again, what other summary statistics could we calculate?
- The horizontal lines on the plot show the sample means for each of the three states.
- In this case, the means appears to capture the magnitude of the data, but what the variability or the spread of the data?
- After all, Florida appears to have more county-to-county variability than the other two states. This can be quantified with the…



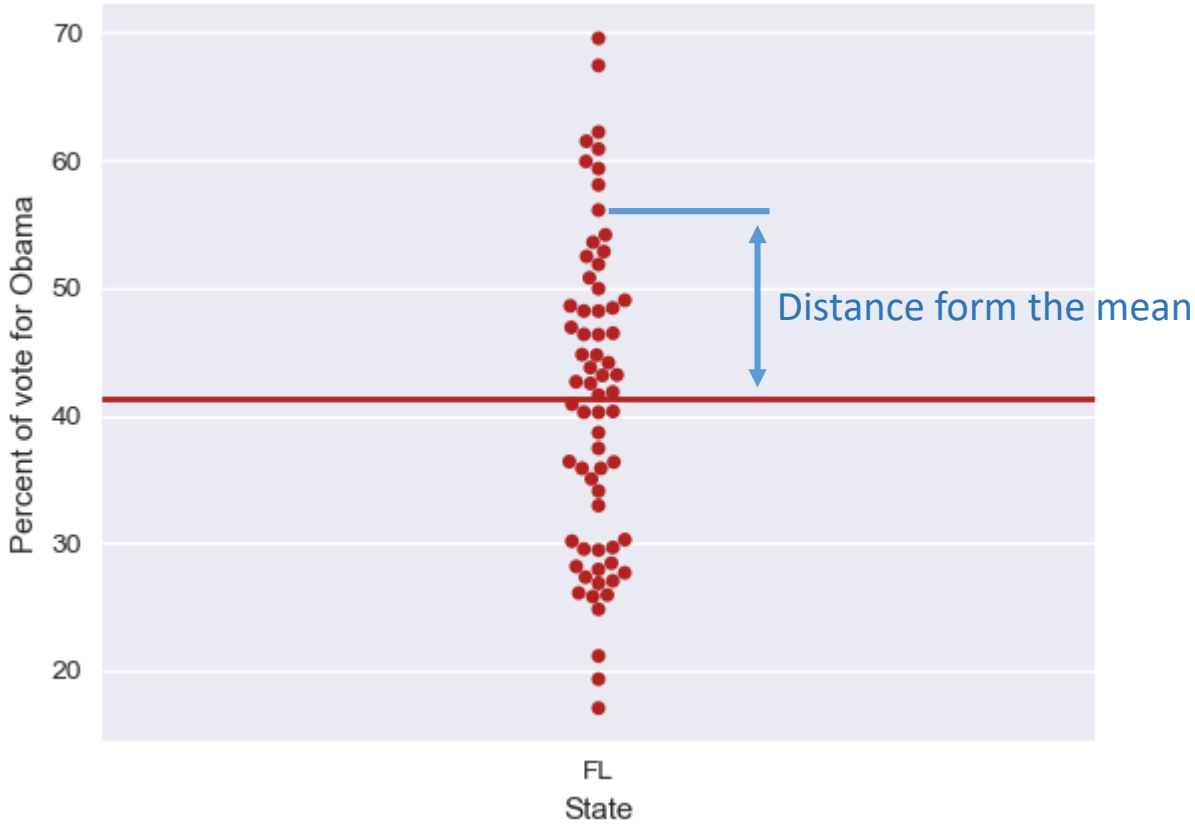
Variance

- The variance is the mean squared distance of the data from their mean.
- In other words, it’s a measure of how far the data are spread.

Calculating the variance

- For each data point, we take the distance from the mean and square it.
- And then take the average of each of these squared values

$$variance = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



Computing the variance

- Calculating the variance in Python is easy:

In:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df_swing = pd.read_csv('C:\Users\lb690\Google Drive\
df_FL = df_swing.loc[df_swing['state'] == 'FL']
FL_mean = np.mean(df_FL['dem_share'])
florida_variance = np.var(df_FL['dem_share'])
print florida_variance
```

Out:

```
147.442786188
```

What is a squared vote?

- The calculation for the variance uses squared values.
- This means that the variance is reported as squared quantities.
- It does not therefore have the same units as to what we measured In this case, in this case, the vote share for Obama. After all, ‘squared votes’, which is not a thing.
- We are therefore interested in the squared root of the variance, which converts the statistic back into its original unit of measurement.
- This is known as…

Standard deviation (σ)

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The standard deviation is calculated in Python with Numpy:

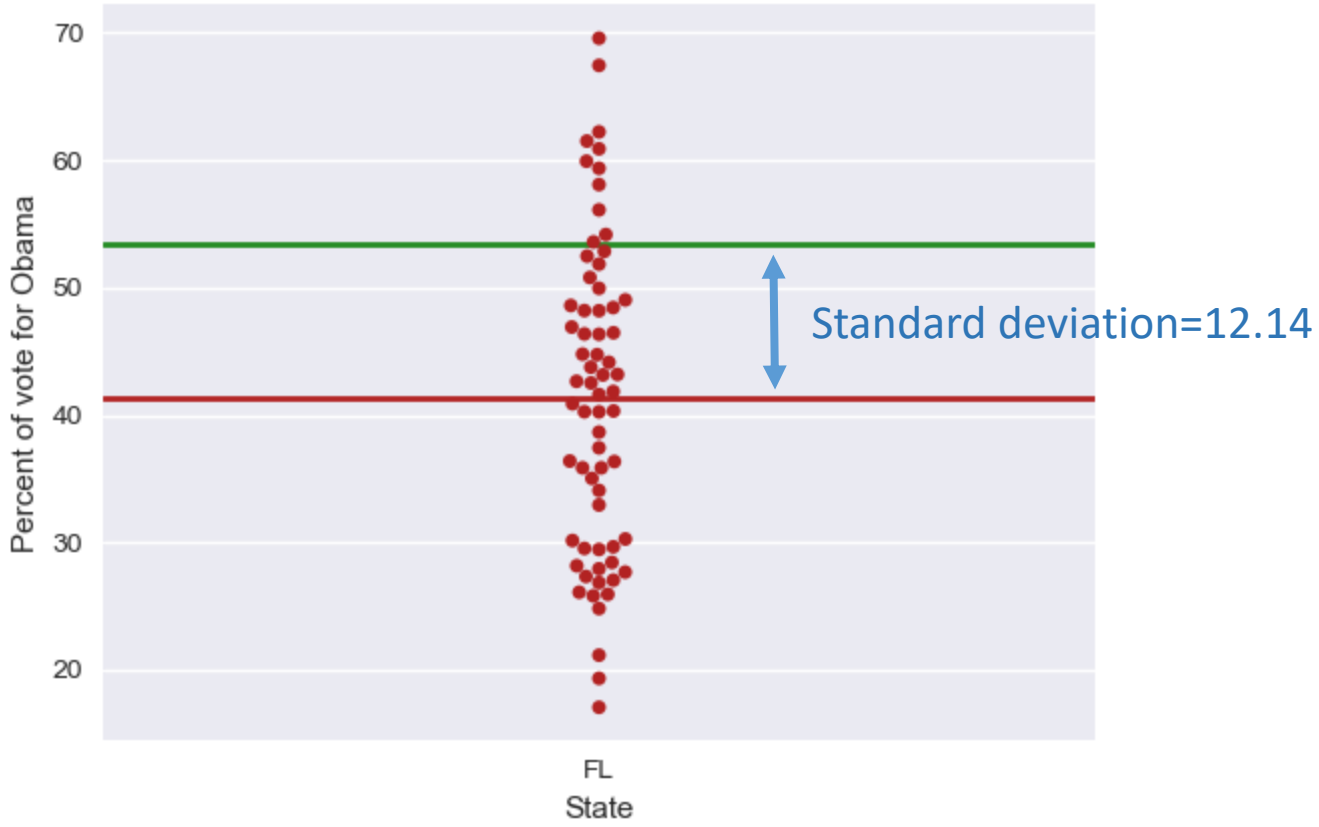
In:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df_swing = pd.read_csv('C:\Users\lb690\Google Drive\Teaching\Stati
df_FL = df_swing.loc[df_swing['state'] == 'FL']
FL_mean = np.mean(df_FL['dem_share'])
florida_std_dev = np.std(df_FL['dem_share'])
print florida_std_dev
```

Out:
```
12.1426021177
```

- You'll notice that, as expected, the answer is the same as that which you would get if you first calculated the variance and then took the square root of the result.

In:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df_swing = pd.read_csv('C:\Users\lb690\Google Drive\
df_FL = df_swing.loc[df_swing['state'] == 'FL']
FL_mean = np.mean(df_FL['dem_share'])
florida_variance = np.var(df_FL['dem_share'])
print "Florida variance:", florida_variance
florida_std_dev = np.sqrt(florida_variance)
print "Florida square root:", florida_std_dev
```

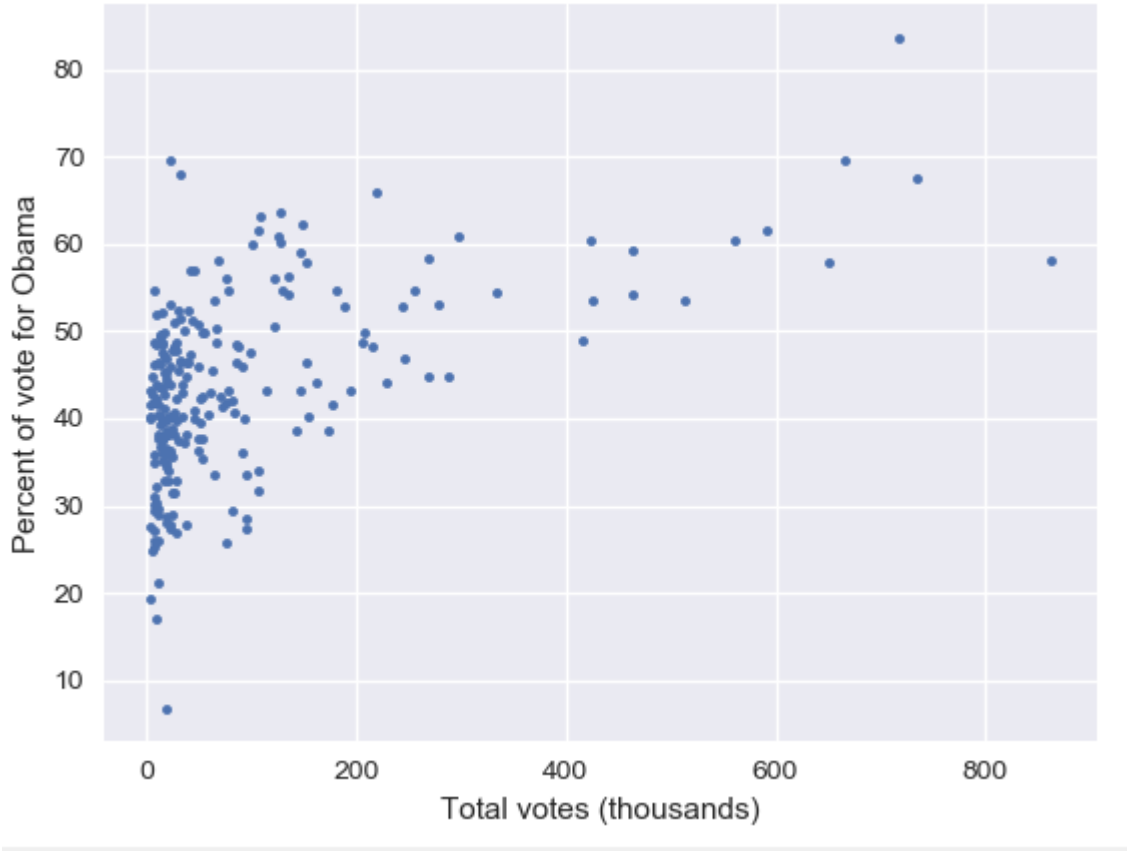
Out:
```
Florida variance: 147.442786188
Florida square root: 12.142602117687158
```


70
60
50
40
30
20
Percent of vote for Obama
FL
State
Standard deviation=12.14

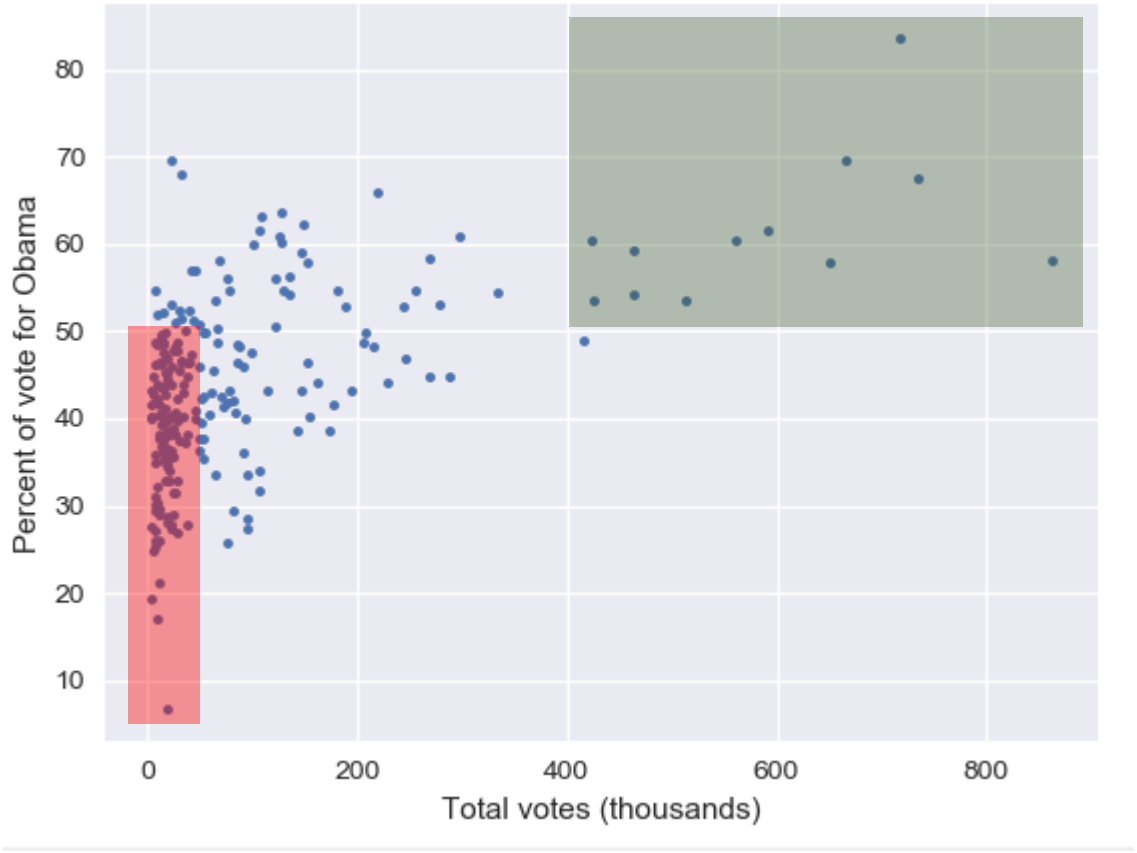
Variation in *one* variable

- So, these four measures all describe aspects of the variation in a single variable:
  - Sum of squared deviations
  - Variance
  - Standard deviation
  - Standard error
- Can we adapt them for thinking about the way in which two variables might vary together?

- Let’s take a look at the total number of votes for each county in the 2008 US election.
- We can start by looking at a scatter plot for the county data for each of the three swing states.



- This immediately shows us that the 12 most populist counties all voted for Obama, and…
- Most of the counties with small populations voted for McCain.

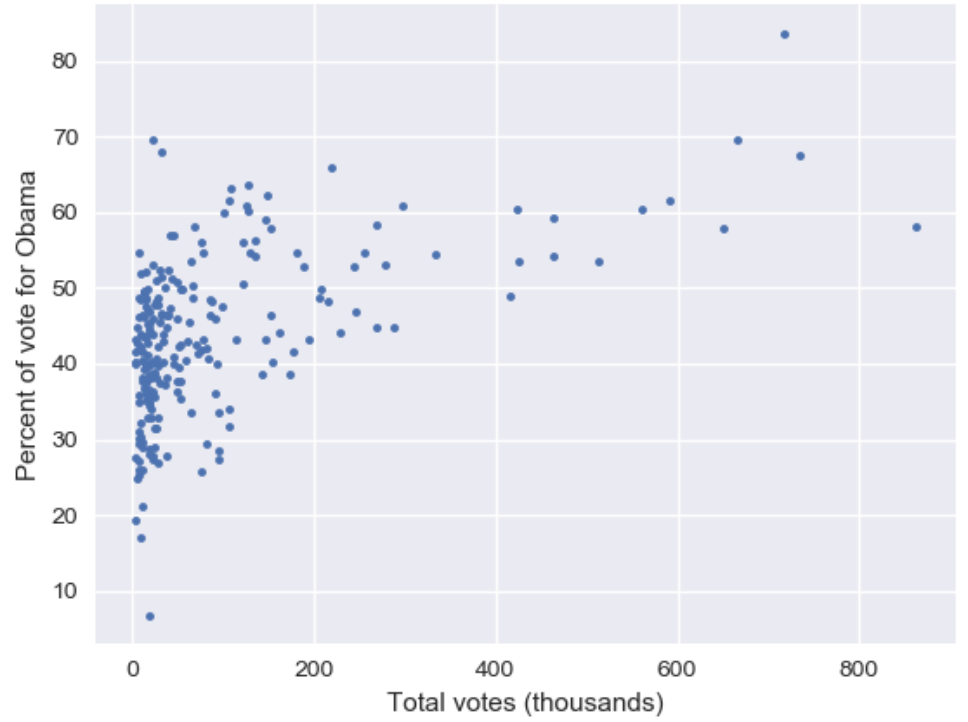


Plotting a scatter plot

In:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df_swing = pd.read_csv('C:\Users\lb690\Google Drive\Teaching\Statistics\Data_and_prep_analysis
_=plt.plot(df_swing['total_votes']/1000, df_swing['dem_share'], marker='.', linestyle='none')
_=plt.xlabel('Total votes (thousands)')
_=plt.ylabel('Percent of vote for Obama')
plt.show()
```

Out:



- It would be good to have a summary statistics to accompany the information that we just got from the scatterplot.
- We want a number the signifies how Obama’s vote share varies with the total vote count.

Covariance

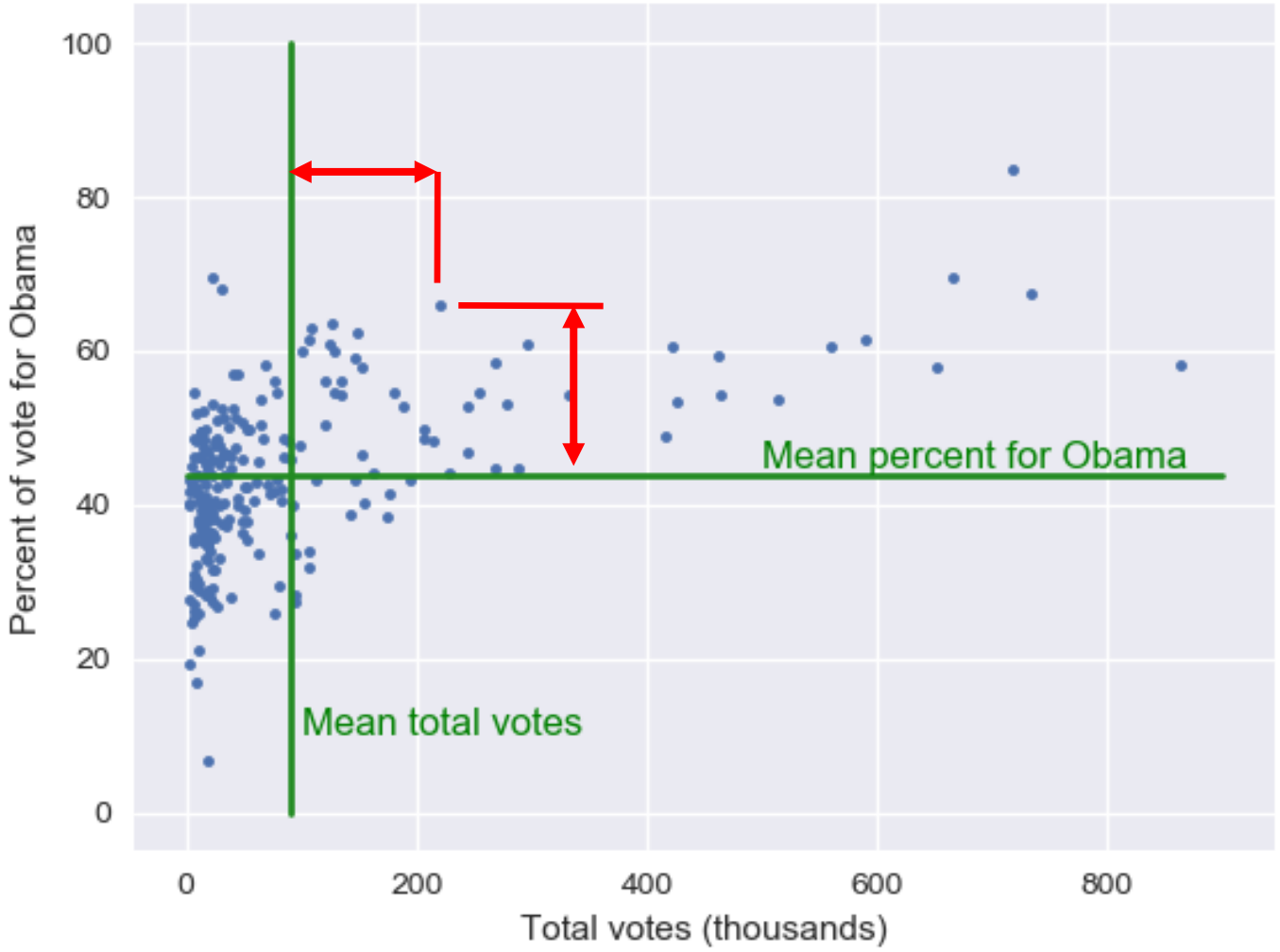
- Covariance is a measure of how two quantities vary together.
- Covariance has some of the properties we want; positive, negative, and absent relationships can be recognised

Calculating covariance

- Let's take a look at this data point, which is Luca County, Ohio.
- It differs from both of the means.
- We can compute these differences for each data point.
- The covariance is then the mean of the product of these differences.

$$covariance = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- However, if we want to have a generally applicable measure, we need it to be dimensionless; to not have any units.



Correlation coefficient - r

- So, we divide the covariance by the standard deviation of the *X* and *Y* variables.
- This is what the product-moment coefficient of correlation (sometimes known as Pearson's coefficient) is.

$$r = \frac{covariance}{(std\ of\ x)(std\ of\ y)}$$

- This is the comparison of the variability in the data due to codependence (covariance) compared to the variability inherent in each of the variables independently (their standard deviations).
- It is completely dimensionless.

The formula has been devised to ensure that the value of r will always lie in the in the range -1 to 1. For example:

r = -1 means a perfect negative correlation

r = 1 means a perfect positive correlation

r = 0 means that there is zero correlation

r = -0.84 means a strong negative correlation

r = 0.15 means a weak positive correlation

And so on…

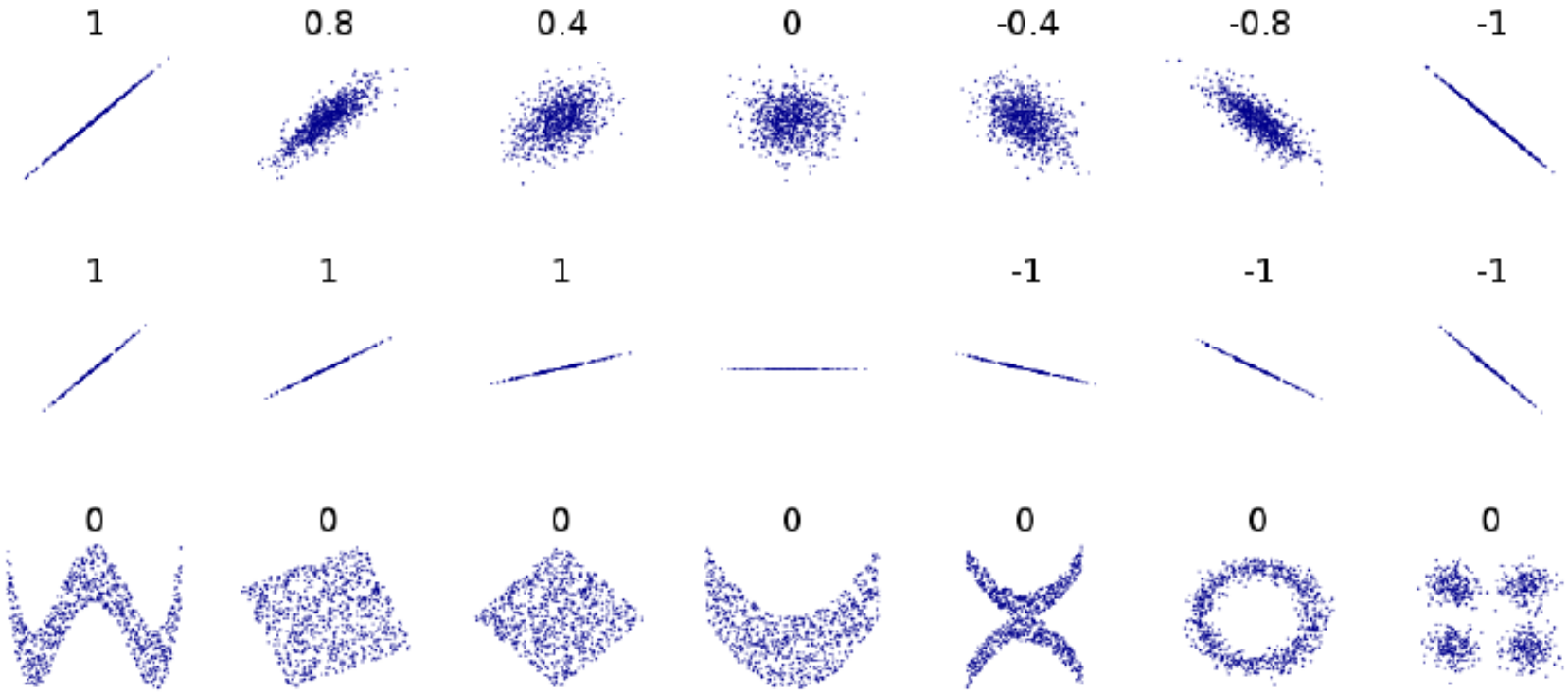
- 0.0 - 0.3: Weak relationship; may be an artefact of the data set and in fact there is no relationship at all.
- 0.3 - 0.6: Moderate relationship; you might be on to something, or you might not.
- 0.6 - 0.9: Strong relationship; you can be confident that these two variables are connected in some way.
- 0.9 - 1.0: Very strong relationship; variables are almost measuring the same thing.

In other words…

- The strength of a correlation will be indicated by how far from zero the value of r lies; with whether it is a positive or negative value indicating whether is it a positive or negative relationship.

Important things to remember

- In general, the smaller the batch size, the larger the value of r has to be in order to provide evidence of significant correlation.
- r only measures linear relationships:



- Correlation is **NOT** causation!

  - Just because *X* and *Y* are correlated does not mean that *X* causes *Y*.

  - They could both be cause by some other factor, *Z*.

  - Or *Y* may cause *X*.

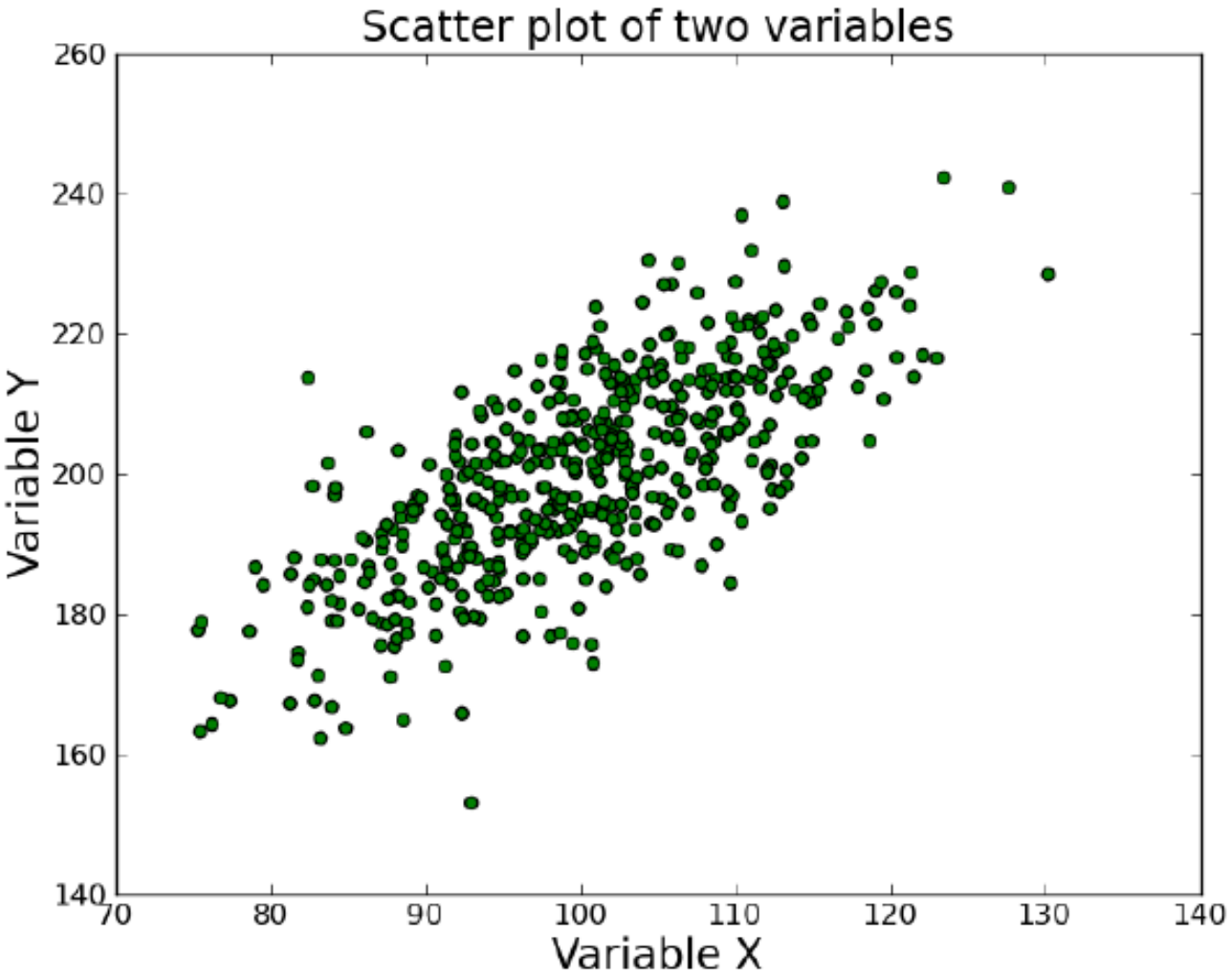
- It is also important to remember that low correlations may be the result of sampling noise, with no causal linkage being present.

Range effects

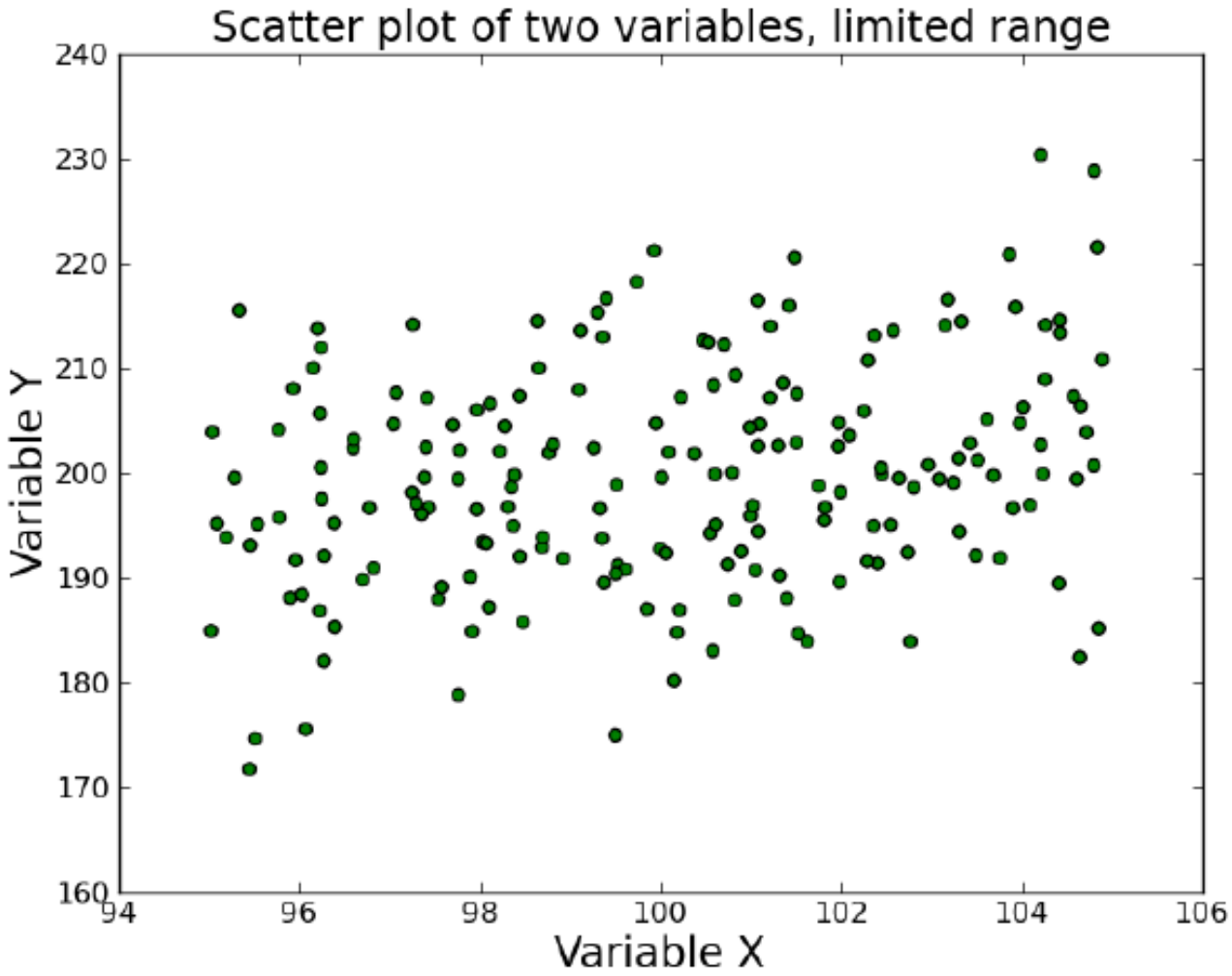
- Two variables can be strongly related across the whole of their range, but with no strong relationship in a limited subset of that range.
- For example, consider the relationship between price and top speed in cars; a broadly positive relationship. However, if we only look at expensive cars, the two values may be uncorrelated.

An example

- Here, *X* is normally distributed, with a mean = 100 and SD = 10.
- For each of the 500 cases, *Y* is equal to *X* plus a normal variate, with mean = 100 and SD = 10.
- *Y* and *X* are clearly related, but there's also a significant part of the variation in *Y* that has nothing to do with *X*.
- In this example, $r = 0.72$.



- However, if we limit the range of *X* to between 95 and 105, the correlation coefficient is only $r = 0.27$.



Information about *Y* from *X*

- If I know the correlation between two things, what does knowing one thing tell me about the value of the other?
- Consider the *X*, *Y* example. *X* was a random variable and *Y* was equal to *X* plus another random variable from the same distribution.
- The correlation worked out at about 0.7. Why?

R-squared

- Turns out that if we square the correlation coefficient, we get a direct measure of the proportion of the variance explained.
- In our example case, we know that *X* explains exactly 50% of the variance in *Y*.
- The square root of $0.5 \approx 0.71$.

- $r = 0.3$ explains 9% of the variance.
- $r = 0.6$ explains 36% of the variance.
- $r = 0.9$ explains 81% of the variance.
- R^2 is a standard way of measuring the proportion of variance we can explain in one variable using one or more other variables.
- This concept links into the analysis of variance (ANOVA).

Any questions?